Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs

Alexey Ozerov, Pierrick Philippe, Frédéric Bimbot, Rémi Gribonval

Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs

Alexey Ozerov, Pierrick Philippe, Frédéric Bimbot, and Rémi Gribonval

***Abstract*—Probabilistic approaches can offer satisfactory solutions to source separation with a single channel, provided that the models of the sources match accurately the statistical properties of the mixed signals. However, it is not always possible to train such models. To overcome this problem, we propose to resort to an adaptation scheme for adjusting the source models with respect to the actual properties of the signals observed in the mix. In this paper, we introduce a general formalism for source model adaptation which is expressed in the framework of Bayesian models. Particular cases of the proposed approach are then investigated experimentally on the problem of separating voice from music in popular songs. The obtained results show that an adaptation scheme can improve consistently and significantly the separation performance in comparison with nonadapted models.**

***Index Terms*—Adaptive Wiener filtering, Bayesian model, expectation maximization (EM), Gaussian mixture model (GMM), maximum *a posteriori* (MAP), model adaptation, single-channel source separation, time–frequency masking.**

I. Introduction

THIS PAPER deals with the general problem of source separation with a single channel, which can be formulated as follows. Let $s_1(n)$ and $s_2(n)$ be two sampled audio signals (also called *sources*) and $x(n)$ the sum of these two signals

$$x(n) = s_1(n) + s_2(n) \tag{1}$$

also called *mix*. Given $x(n)$, the source separation problem in the case of a single channel consists in estimating the contributions $\hat{s}_k(n)$ of each of the two sources $(k = 1, 2)$.

Several methods (for example [1]–[4]) have been proposed in the literature to approach this problem. In this paper, we consider the probabilistic framework, with a particular focus on Gaussian mixture models (GMMs) [5], [6]. The GMM-based approach offers the advantage of being sufficiently general and applicable to a wide variety of audio signals. These methods have indeed shown good results for the separation of speech signals [5] and some particular musical instruments [6].

The underlying idea behind these techniques is to represent each source s_k by a GMM, which is composed by a set of characteristic spectral patterns. Each GMM is learned on a training set, which contains samples of the corresponding audio class (for instance, speech, music, drums, etc.). In this paper, we refer to these models as *general* or *a priori* models, as they are supposed to cover the range of properties observable for sources belonging to the corresponding class.

An efficient model must be able to yield a rather accurate description of a given source or class of sources, in terms of a collection of spectral shapes corresponding to the various behaviors that can be observed in the source realizations. This requires GMMs with a large number of Gaussian functions, which raises a number of problems:

- *trainability* issues linked to the difficulty in gathering and handling a representative set of examples for the sources or classes of sources involved in the mix;
- *selectivity* issues arising from the fact that the particular sources in the mix may only span a small range of observations within the overall possibilities covered by the general models;
- sensor and channel *variability* which may affect to a large extent the acoustic observations in the mix and cause a more or less important mismatch with the training conditions;
- computational *complexity* which can become intractable with large source models, as the separation process requires factorial models [5], [6].

A typical situation which illustrates these difficulties arises for the separation of voice from music in popular songs. For such a task, it turns out to be particularly unrealistic to accurately model the entire population of music sounds with a tractable and efficient GMM. The problem is all the more acute as the actual realizations of music sounds within a given song cover much less acoustic diversity than the general population of music sounds.

The approach proposed in this paper is to resort to model *adaptation* in order to overcome the aforementioned difficulties. In a similar way, as it is done for instance in speaker (or channel) adaptation for speech recognition, the proposed scheme consists in adjusting the source models to their realizations in the mix $x(n)$. This process intends to specialize the *adapted* or *a posteriori* models to the particular properties of the sources *as observed in the mix*, while keeping the model complexity tractable.

Manuscript received July 12, 2006; revised March 6, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Te-Won Lee.

A. Ozerov was with Orange Labs, 35512 Cesson Sévigné Cedex, France, and IRISA (CNRS and INRIA), Metiss Group (Speech and Audio Processing), 35042 Rennes Cedex, France. He is now with the Sound and Image Processing (SIP) Laboratory, KTH (Royal Institute of Technology), SE-100 44 Stockholm, Sweden (e-mail: loha_ozerov@mail.ru).

P. Philippe is with Orange Labs, 35512 Cesson Sévigné Cedex, France (e-mail: pierrick.philippe@orange-ftgroup.com).

F. Bimbot and R. Gribonval are with IRISA (CNRS and INRIA), Metiss Group (Speech and Audio Processing), 35042 Rennes Cedex, France (e-mail: frederic.bimbot@irisa.fr; remi.gribonval@irisa.fr).

Digital Object Identifier 10.1109/TASL.2007.899291

In the first part of this article, we propose a general formalism for model adaptation in the case of mixed sources. This formalism is founded on Bayesian modeling and statistical estimation with missing data.

The second part of the work is dedicated to experiments and assessment of the proposed approach in the case of voice/music separation in popular songs. We show how separation performance can be significantly improved with model adaptation.

The remainder of the paper is structured as follows. In Section II, the principles of probabilistic single-channel source separation are presented, the limitations of this approach are discussed and the problem studied in this paper is defined. Then, in Section III, a general formalism for source model adaptation is presented and further developed in the particular case of a maximum *a posteriori* (MAP) criterion. Section IV is dedicated to the customization of the proposed approach to the problem of voice/music separation in monophonic popular songs. Finally, Section V presents the experimental results, with simulations and evaluations which validate the proposed approach. All technical aspects of the paper, including the precise description of the adaptation algorithms, are gathered in an Appendix.

II. Probabilistic Single-Channel Source Separation

A. Source Separation Based on Probabilistic Models: General Framework

The problem of source separation with a single channel, as formulated in (1), is fundamentally ill-posed. In fact, for any signal $z(n)$, the couple $(\hat{s}_1(n) = z(n), \hat{s}_2(n) = x(n) - z(n))$ is a solution to the problem.

Therefore, it is necessary to express additional constraints or hypotheses to elicit a unique solution. In the case of the probabilistic approach, the sources s_1 and s_2 are supposed to have a different statistical behavior, corresponding to different known source models Λ_k, $k = 1, 2$ with $\Lambda_1 \neq \Lambda_2$. Therefore, among all possible solutions to (1), one can choose the pair minimizing some distortion measure given these models. This can be expressed as the optimization of the following criterion, subject to the constraint (1):

$$\min_{\hat{s}_1, \hat{s}_2} \mathbb{E}_{s_1, s_2} \left[d(\hat{s}_1, \hat{s}_2; s_1, s_2) | x, \Lambda_1, \Lambda_2 \right] \tag{2}$$

where $d(\cdot)$ is a distortion measure between the sources s_k and their estimates $\hat{s}_k$. Since the sources are not observed, the value of this function is replaced by its expectation conditionally on the observed mix and the source models.

The source models are generally trained on databases of examples of audio signals, the characteristics of which are close to those of the sources within the mix [5], [7]. In this paper, such models will be referred to as *general* models.

The separation problem of (1) can be reformulated in the short-time Fourier transform (STFT) domain [5], [6]. Since the STFT is a linear transform, we have

$$X(t, f) = S_1(t, f) + S_2(t, f) \tag{3}$$

where $X(t, f)$, $S_1(t, f)$ and $S_2(t, f)$ denote the STFT of the time-domain signals $x(n)$, $s_1(n)$, and $s_2(n)$ for each signal frame number $t = 1, 2, \ldots, T$ and each frequency index $f = 1, 2, \ldots, F$ (F being the index of the Nyquist frequency). In the rest of the paper, the presentation will take place in the STFT domain, knowing that the OLA (Overlap and Add) method can be used to reconstruct the signal (see for instance [8]). Time-domain signals will be systematically denoted by a lower-case letter, while STFT-domain quantities will be denoted by their upper-case counterpart.

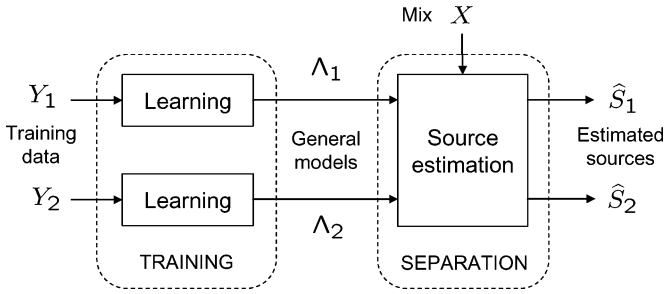

Fig. 1. Source separation based on general *a priori* probabilistic models.

The formulation of the problem in the STFT domain is motivated by the fact that audio sources are generally weakly overlapping in the time–frequency domain. This property has been illustrated for instance in [9]. In fact, if the sources do not overlap at all in the STFT-domain, i.e., if $S_1(t, f)S_2(t, f) = 0$ for any t and f, the following masking operation yields the exact solution for the estimation of the kth source:

$$\widehat{S}_k(t, f) = \mathcal{M}_k(t, f)X(t, f) \tag{4}$$

where $\mathcal{M}_k(t, f) = 1$, if $S_k(t, f) \neq 0$, and $\mathcal{M}_k(t, f) = 0$, otherwise.

As, in practice, the sources overlap partly, this approach can be adjusted by using a masking function (or *mask*) that takes continuous values $\mathcal{M}_k(t, f) \in [0, 1]$ and choosing $\mathcal{M}_k(t, f)$ close to 1 if the kth source is dominant in the time–frequency region defined by (t, f) and close to 0 if the kth source is dominated. In that case, the masking approach does not yield the exact solution, but an optimal one in some weighted least-square sense. The operation expressed in (4) is called *time–frequency masking*, and it also corresponds to an adaptive filtering process.

However, the main difficulty is that the knowledge on the respective dominance of the sources within the mix is not available, which makes it impossible to obtain an exact estimation of the optimal masks $\mathcal{M}_k = \{\mathcal{M}_k(t, f)\}_{t,f}$. In the conventional probabilistic approach, the source models are used to estimate masking functions according to the observed behavior of the sources.

Fig. 1 summarizes the general principles of probabilistic source separation. The general models Λ_1 and Λ_2 are trained independently on sets of examples Y_1 and Y_2. The source estimates $\widehat{S}_1$ and $\widehat{S}_2$ are obtained by filtering the mix X (cf. (4)) with masks estimated from the general source models Λ_1 and Λ_2 and the mix itself X.

1) GMM Source Model: As mentioned earlier, the approach reported on in this article is based on Gaussian mixture models (GMMs) of the audio sources. A number of recent works have been using GMMs or, more generally, hidden Markov models

(HMMs) to account for the statistical properties of audio sources [5]–[7], [10]–[13], the latter being a rather natural extension of the former. The GMM/HMM-based framework allows to model and to separate nonstationary audio sources, as considered here, assuming that each source is locally stationary and modeled by a particular Gaussian within the corresponding mixture of Gaussians.

The underlying idea is to represent each source as the realization of a random variable driven by a finite set of characteristic spectral shapes, i.e., "local" power spectral densities (PSDs). Each local PSD describes some particular sound event. Under the GMM formalism, model Λ_k for the kth audio source is composed of Q_k states corresponding to Q_k local PSDs $\{r^2_{k,i}(f)\}_{1\leq f\leq F}$, $i=1,2,\ldots,Q_k$.

Conditionally, to state i, the short-term spectrum $S_k(t)$ is viewed as some realization of a random Gaussian complex vector with zero mean and diagonal covariance matrix $R_{k,i}$ corresponding to the local PSD, i.e., $R_{k,i}=\mathrm{diag}[\{r^2_{k,i}(f)\}_f]$. Such a GMM can be parameterized as $\Lambda_k=\{u_{k,i},R_{k,i}\}_i$, where $u_{k,i}\geq 0$ are the weights of each Gaussian density satisfying $\sum_i u_{k,i}=1$. Altogether, the GMM probability density function (pdf) of the short-term spectrum $S_k(t)$ can be written as

$$p\left(S_k(t)|\Lambda_k\right)=\sum_i u_{k,i}N_C\left(S_k(t);\bar{0},R_{k,i}\right) \tag{5}$$

where $N_C(V;\mu,R)$ denotes the pdf of a *complex* Gaussian random vector $V\in\mathbb{C}^F$ with mean vector $\mu=\{\mu(f)\}_f\in\mathbb{C}^F$ and diagonal covariance matrix $R=\mathrm{diag}[\{r^2(f)\}_f]\in\mathbb{R}^{F\times F}$, defined as in [14] (pp. 503–504)

$$N_C(V;\mu,R)=\prod_f\frac{1}{\pi r^2(f)}\exp\left[-\frac{|V(f)-\mu(f)|^2}{r^2(f)}\right]. \tag{6}$$

2) Model Learning: A conventional framework for learning the GMM's parameters Λ_k, $k=1,2$ from training data Y_k (for kth source) is based on optimizing the maximum-likelihood (ML) criterion

$$\Lambda_k=\arg\max_{\Lambda'_k}p\left(Y_k|\Lambda'_k\right). \tag{7}$$

This approach is used for source separation, for instance, in [5]–[7]. In practice, the optimization of the ML criterion is obtained with an expectation-maximization (EM) algorithm [15].

3) Source Estimation: Once the source models trained, the sources in the mix can be estimated in the minimum mean square error (mmse) sense, i.e., with the distortion measure from (2) defined as $d(\hat{s}_1,\hat{s}_2;s_1,s_2)=\|S_1-\widehat{S}_1\|^2+\|S_2-\widehat{S}_2\|^2$. This leads to a variant of adaptive Wiener filtering, which is equivalent to the time–frequency masking operation (4) with the mask $\mathcal{M}_1$ being calculated as follows [6] (and similarly for $\mathcal{M}_2$):

$$\mathcal{M}_1(t,f)=\sum_{i,j}\gamma_{i,j}(t)\frac{r^2_{1,i}(f)}{r^2_{1,i}(f)+r^2_{2,j}(f)} \tag{8}$$

where $\gamma_{i,j}(t)$ denotes the *a posteriori* probability that the state pair (i,j) has emitted the frame t, with the property that $\sum_{i,j}\gamma_{i,j}(t)=1$, and

$$\begin{aligned}\gamma_{i,j}(t)&\triangleq P\left(q_1(t)=i,q_2(t)=j|X,\Lambda_1,\Lambda_2\right)\\&\propto u_{1,i}u_{2,j}N_C\left(X(t);\bar{0},R_{1,i}+R_{2,j}\right)\end{aligned} \tag{9}$$

where the symbol $\propto$ denotes proportionality, $X(t)$ the short-term spectrum of the mix, and $q_1(t)$, $q_2(t)$ the hidden states in models Λ_1 and Λ_2.

B. Problem Statement

In the approach presented in the previous subsections, a difficulty arises in practice from the fact that the source models Λ_k tend to perform poorly in realistic cases, as there is generally a mismatch between the models and the actual properties of the sources in the mix.

To illustrate this issue, let us take an example where one of the sources is a voice signal (as, for instance in [5], [7], [10], and [12]). Either the voice model has been trained on a particular voice but its generalization ability tends to be poor to other voices, or it is trained on a group of voices but then it requires a large number of parameters, and even though, it tends to lack selectivity to a particular voice in a particular mix, not to mention the variability problems that can be caused by different recording and/or transmission conditions.

The same problem is reported also with other classes of signals, in particular, musical instruments [6], [11], all the more acute as the separation problem is formulated with less *a priori* knowledge, for instance, separating singing voice from music, where the class of music signals is extremely wide.

Thus, the practical use of statistical approaches to source separation requires the following problems to be addressed:

1) deal with the scarcity of representative training data for wide classes of audio signals (for instance, the class of music sounds);
2) specialize a source model to the particular properties of a given source in a given mix (for instance, a particular instrument or combination of instruments);
3) account for recording and transmission variability which can affect significantly the statistical behavior of a source in a mix, w.r.t. its observed properties in a training data set (for instance, the type of microphone, the room acoustics, the channel distortion, etc.);
4) control the computational complexity which arises when dealing with large-size statistical models (for instance, hundreds or thousands of Gaussian functions in a GMM).

These problems can be formulated more strictly in terms of statistical modeling. Suppose that the source S_k observed in the mix is the realization of a random process $\mathcal{S}_k$, and that the training data Y_k is the realization of a (more or less slightly) different random process $\mathcal{Y}_k$. Let us denote as $p_{\mathcal{S}_k}(\cdot)$ and $p_{\mathcal{Y}_k}(\cdot)$ the pdfs of these two processes.

In order to reliably estimate S_k, $k=1,2$ (Section II-A3), the ideal situation would be to know their exact pdfs $p_{\mathcal{S}_k}(\cdot)$.

However, the sources are not observed separately, which makes it impossible to access or even estimate reliably their pdfs. These pdfs are therefore replaced by those of the training data $p_{\mathcal{Y}_k}(\cdot)$ and approximated by GMMs $p(\cdot|\Lambda_k)$ optimized according to the ML criterion as in (7). In summary

$$p_{\mathcal{S}_k}(\cdot) \approx p_{\mathcal{Y}_k}(\cdot) \approx p(\cdot|\Lambda_k). \tag{10}$$

Model learning with a training scheme requires the training data to be extremely representative of the actual source properties in the mix, which means very large databases with high coverage. However, the effective use of the models for source separation implies that they are also rather selective, i.e., well-fitted to the actual statistical properties of each source in the mix.

In order to overcome these limitations, we propose to resort, when possible, to an *adaptation* scheme which aims at adjusting *a posteriori* the models by tuning their characteristics to those of the sources actually observed in the mix. As it will be detailed further, this approach makes it possible, under certain conditions, to improve the quality of the source model, while keeping its dimensionality reasonable.

III. Model Adaptation With Missing Acoustic Data

The goal of model adaptation is to replace the general models (which match well the properties of the training sources, but not necessarily those of the corresponding sources in the mix), with *adapted models* adjusted so as to better represent the sources in the mix, thus leading to an improved separation ability. In this section, model adaptation is introduced in a general form. The principle is then detailed in the case of a MAP adaptation criterion.

A. Principle

In contrast with the general models Λ_1 and Λ_2, *adapted* models have their characteristics tuned to those of the sources in the mix. Although, adapted and general models have exactly the same structure, new notations are introduced for adapted models and for their parameters, in order to distinguish between these two types of models. Thus, the adapted models are denoted λ_k, $k = 1, 2$ and parameterized as $\lambda_k = \{\omega_{k,i}, \Sigma_{k,i}\}_i$ with $\omega_{k,i}$ being weights of Gaussians and $\Sigma_{k,i} = \mathrm{diag}[\{\sigma^2_{k,i}(f)\}_f]$ covariance matrices.

The ideal situation for model adaptation would be to learn the models from the test data, i.e., from the separated sources S_k, or at least from some other sources $\tilde{S}_k$ having characteristic extremely similar to those of S_k. For example, Benaroya *et al.* [6] evaluate their algorithms in such a context. They learn the models from the separated sources (available in experimental conditions) issued from the first part of a musical piece, and then they separate the second part of the same piece. While the results are convincing, such a procedure is only possible in a rather artificial context.

Another interesting direction is to try to infer the model parameters directly from the mix X. For example, Attias [16] uses such an approach in the multichannel *determined* case, when there are at least as many channels (or mixes) as sources. In this case, the spatial diversity (i.e., the fact that the sources come from different directions) creates a situation which allows to estimate the models without any other *a priori* knowledge. In the single-channel case studied here, this approach cannot be applied as it is, since the spatial diversity is not exploitable. Indeed, one could try to look for the models λ_1 and λ_2 optimizing the following ML criterion:

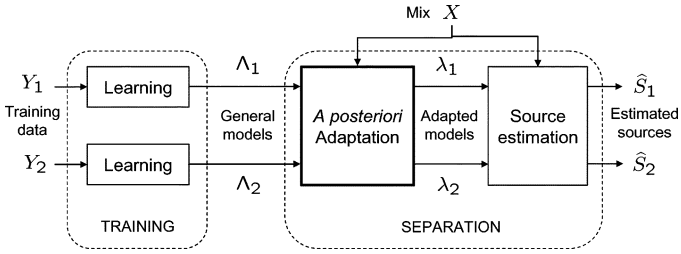

Fig. 2. Source separation based on adapted *a posteriori* probabilistic models.

$$\left(\lambda_1^{\mathrm{ML}}, \lambda_2^{\mathrm{ML}}\right) = \arg \max_{(\lambda_1', \lambda_2')} p\left(X|\lambda_1', \lambda_2'\right) \tag{11}$$

but this would certainly not lead to any good model estimates, since in this criterion there is no *a priori* knowledge about the sources to distinguish between them. For example, swapping the models λ_1 and λ_2 in this criterion does not change the value of the likelihood, i.e., $p(X|\lambda_1, \lambda_2) = p(X|\lambda_2, \lambda_1)$.

An alternative approach is to use the MAP adaptation approach [17] widely applied for speech recognition [18] and speaker verification [19] tasks. The MAP estimation criterion consists in maximizing the *posterior* $p(\lambda_1, \lambda_2|X)$ rather than the likelihood $p(X|\lambda_1, \lambda_2)$, as in (11). Using the Bayes rule, this *posterior* can be represented as $p(\lambda_1, \lambda_2|X) \propto p(X|\lambda_1, \lambda_2)p(\lambda_1, \lambda_2)$ with a proportionality factor which does not depend on the models λ_k, $k = 1, 2$ and therefore has no influence on the optimization of the criterion. In contrast to the ML criterion, the model parameters are now considered as realizations of some random variables and their *a priori* (or *prior*) pdf $p(\lambda_1, \lambda_2)$ should be specified. We suppose that the parameters of model λ_1 are independent from those of model λ_2 and that the pdf of the parameters of each model depends on the parameters of the corresponding general model, which can be summarized as $p(\lambda_1, \lambda_2) \triangleq p(\lambda_1|\Lambda_1)p(\lambda_2|\Lambda_2)$. Finally, we have the following MAP criterion:

$$\begin{aligned}\left(\lambda_1^{\mathrm{MAP}}, \lambda_2^{\mathrm{MAP}}\right) \\ = \arg \max_{(\lambda_1', \lambda_2')} p\left(X|\lambda_1', \lambda_{2'}\right) p\left(\lambda_1'|\Lambda_1\right) p\left(\lambda_2'|\Lambda_2\right)\end{aligned} \tag{12}$$

Note that the MAP criterion (12), in contrast to the ML criterion (11), involves the prior pdfs $p(\lambda_k|\Lambda_k)$, $k = 1, 2$ which forces the adapted models to stay attached to the general ones. Thus, the general models play the role of *a priori* knowledge about the sources.

Better separation performance may be achieved with the MAP criterion (12) and appropriate priors compared to what can be obtained with general models. Fig. 2 illustrates the integration of the *a posteriori* adaptation unit into the baseline separation scheme (Fig. 1).

For the sake of generality, we do not give here any functional form for the prior pdfs $p(\lambda_k|\Lambda_k)$. A discussion concerning the role of the priors is proposed in Section III-A1, together

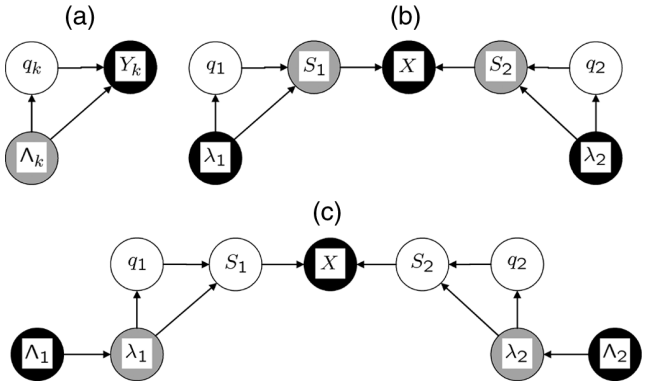

Fig. 3. Bayesian networks representing model learning, source estimation, and *a posteriori* model adaptation (Fig. 2). Recall that $q_k = \{q_k(t)\}_t$, $k = 1, 2$ denote GMM state sequences. Shadings of nodes: observed nodes (black), estimated hidden nodes (gray), and other hidden nodes (white). (a) Learning. (b) Source estimation. (c) *A posteriori* model adaptation.

with some ideas on how to choose them. In Section IV-D1, these priors are represented as parametric constraints, thus introducing a class of *constrained adaptation* techniques. Two particular adaptation techniques (i.e., filter and PSD gains adaptation) belonging to this class are introduced in Sections IV-D2 and IV-D3 and evaluated in the experimental part of this paper.

The general adaptation scheme can be represented as in Fig. 3 using Bayesian networks (or oriented graphical models) [20], [21] in order to give some graphical interpretation of the dependencies. Different shadings are used in order to distinguish between different types of nodes. Observed nodes are in black, hidden nodes estimated conditionally on the observed nodes are in gray, and all other hidden nodes (nonestimated ones) are in white.

We propose to call the approach presented in this article *model adaptation with missing acoustic data*. This expression reflects two following ideas.

1) *Model adaptation* corresponds to the attachment of the adapted models to the general models, for instance by means of prior pdfs $p(\lambda_k|\Lambda_k)$, as in (12).
2) The use of *missing acoustic data* corresponds to the fact that the model parameters are estimated from the mix X, whereas the actual acoustic data (the sources S_k) are unknown (i.e., missing). The adjective *acoustic* is added in order to avoid any confusion with *missing data* from the EM algorithm's terminology [15].

1) Role of Priors in the MAP Approach: In the case of the MAP approach, the choice of the prior pdfs $p(\lambda_k|\Lambda_k)$ results from a tradeoff. On the one hand, since the adaptation is carried out from the mix X, the priors should be restrictive enough to attach well the adapted models λ_k to the general ones Λ_k. On the other hand, the priors should still give some freedom to models, so that they can be adapted to the characteristics of the mixed sources. Two extreme cases of this tradeoff are as follows.

1) The models λ_k could be completely attached to the general models Λ_k, i.e., there is no adaptation freedom and $\lambda_k = \Lambda_k$. This is equivalent to the separation scheme without adaptation (Fig. 1).
2) The adapted models λ_k could be completely free, i.e., the priors are noninformative uniform ($p(\lambda_k|\Lambda_k) \propto const$). This is the case of the ML criterion (11) which, as already discussed, may not lead to a satisfactory adaptation.

A good choice of the priors is therefore crucial, and some examples of potentially applicable priors could be inspired by many adaptation techniques used for speech recognition and speaker verification, such as MAP [17], [19], maximum likelihood linear regression (MLLR) [22], [23], structural MAP (SMAP)[4], eigenspace-based MLLR (EMLLR) [25], etc. Lee and Huo [18] propose a review of all these methods.

Note that the MAP adaptation as presented in [17] and [19] corresponds to a particular choice of conjugate priors (normal inverse Wishart priors for covariance matrices, and Dirichlet priors for Gaussian weights). In this paper, we call *MAP adaptation* any procedure which can be represented in the form of the MAP criterion (12) whatever the priors.

2) Comparison With the State of the Art: For source separation with a single channel, some authors propose to introduce invariance to some physical characteristics into source modeling. For example, Benaroya *et al.* [26] use time-varying gain factors, thus introducing an invariance to the local signal energy. For musical instruments separation, Vincent *et al.* [11] propose to use other descriptive parameters representing the volume, the pitch, and the timbre of the corresponding musical note. These additional parameters are estimated *a posteriori* for each frame, since they are time varying. Thus, these approaches can also be considered as an adaptation process. Note, however, that this type of adaptation is based on the introduction of additional parameters which modify the initial structure of the models.

In order to complete the positioning of our work, it must be underlined that the approach formalized in this article groups two aspects together. The *adaptation* aspect is inspired by the adaptation techniques used for instance for speech recognition and speaker verification tasks [17]–[19], [22]–[25]. The *inference* of model parameters from the mix shares some common points with works concerning speaker identification in noise [27] and blind clustering of popular music recordings [28]. In these two works, the first model is estimated from the mix with the second one fixed *a priori*, but there is no notion of adaptation, i.e., no attachment of the estimated models to some general ones.

Therefore, our article proposes two main contributions. As developed above, it is possible to group in a same formalism the *adaptation* aspect and *inference* of model parameters from the mix. Details on the corresponding algorithms, in the MAP framework, are provided in the Appendix.

The second contribution is the customization, experimentation, and application of this formalism in a particular case of single-channel source separation. This is detailed in the upcoming sections.

IV. Application to Voice/Music Separation in Popular Songs

The proposed formalism concerning model adaptation is further developed in this section, with the purpose to customize it to a particular separation task: the separation of singing voice from music in popular songs.

This separation task is particularly useful for audio indexing. Indeed, the extraction of metadata used for indexing (such as

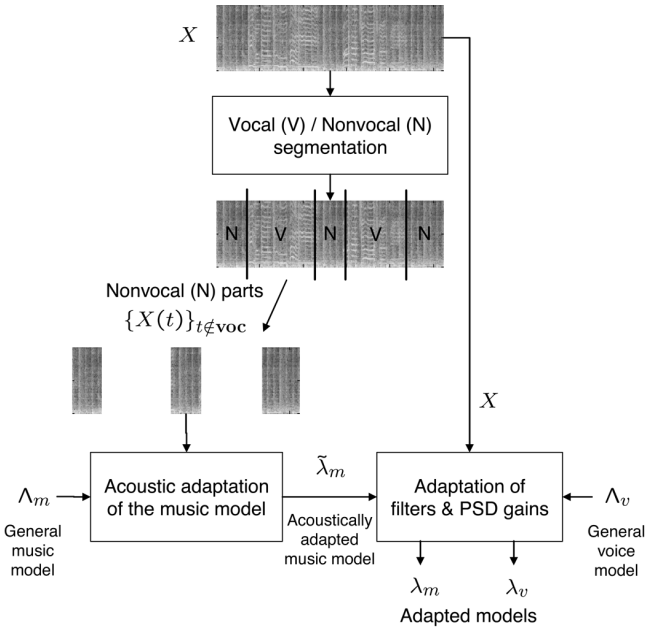

Fig. 4. *A posteriori* model adaptation block (compare to Fig. 2) for voice/music separation.

melody, some keywords, singer identity, etc) is likely to be much easier using separate voice and music signals rather than voice mixed with music.

As in (1), it is assumed that each song's recording $x(n) = s_v(n) + s_m(n)$ is a mix of two sources now denoted $s_v(n)$ (for voice) and $s_m(n)$ (for music). The problem is to estimate the contribution of voice $\hat{s}_v(n)$ and that of music $\hat{s}_m(n)$ given the mix $x(n)$.

For this particular task, the source separation system is designed according to Fig. 2. Model learning and source estimation blocks are implemented as they are described in Sections II-A2 and II-A3. The remainder of this section is devoted to the description of the model adaptation block.

A. Overview of the Model Adaptation Block

In popular songs, there are usually some portions of the signal when music appears alone (free from voice). We call the corresponding temporal segments *nonvocal parts* in contrast to *vocal parts*, i.e., parts that include voice. A key idea in our adaptation scheme is inspired by the work of Tsai *et al.* [28] and is to use the nonvocal parts for music model adaptation. Then, the obtained music model and the general voice model are further adapted on the totality of the song.

The proposed model adaptation block is represented in Fig. 4 and consists of the following three steps.

1) The song X is first segmented into vocal parts $\{X(t)\}_{t\in\mathbf{voc}}$ and nonvocal parts $\{X(t)\}_{t\notin\mathbf{voc}}$ (here $\mathbf{voc}$ denotes the set of vocal frames indices).
2) An acoustically adapted music model $\tilde{\lambda}_m$ is estimated from the nonvocal parts $\{X(t)\}_{t\notin\mathbf{voc}}$ (see Section IV-C).
3) The acoustically adapted music model $\tilde{\lambda}_m$ and the general voice model Λ_v are further adapted on the entire song X with respect to *adaptation of filters and PSD gains* (presented in Section IV-D).

The resulting models are then used to separate the sources according to Fig. 2.

In summary, the music model is first adapted alone so that it better reflects the acoustic characteristics of the very type of music in the song and then both music and voice models are adapted in terms of gain level and recording conditions.

The functional blocks of this adaptation scheme (Fig. 4) are described in the following sections.

B. Automatic Vocal/Nonvocal Segmentation

The practical problem of segmenting popular songs into vocal and nonvocal parts was already studied [29]–[32], and some reported systems give reasonable segmentation performance.

In the work reported in this paper, a classical solution based on GMMs [28], [32] is used. The STFT of processed song $X = \{X(t)\}_t$, which is a sequence of short-time spectra, is transformed into a sequence of acoustic parameters $\mathcal{X} = \{\mathcal{X}(t)\}_t$ (typically MFCCs [33]). Two GMMs Γ_V and Γ_N modeling, respectively, vocal and nonvocal frames are used to decide if the vector $\mathcal{X}(t)$ is a vocal or a nonvocal one.[1] The GMMs Γ_V and Γ_N are learned from some training data, i.e., popular songs manually segmented into vocal and nonvocal parts. These models are used for segmentation without any preliminary adaptation to the characteristics of the processed song. These are indeed general segmentation models.

The vocal/nonvocal decision for the tth frame can be obtained by comparing the log-likelihood ratio with some threshold ψ:

$$\log p\left(\mathcal{X}(t)|\Gamma_V\right) - \log p\left(\mathcal{X}(t)|\Gamma_N\right) \underset{\text{nonvoc}}{\overset{\text{voc}}{\gtrless}} \psi. \qquad (13)$$

However, the segmentation performance can be increased significantly by averaging the frame-based score over a block of several consecutive frames [28], [32]. For this *block-based decision*, the log-likelihood ratio (13) over each block of $U = 2L+1$ frames is computed as

$$\frac{1}{U}\sum_{l=t-L}^{t+L}\left[\log p\left(\mathcal{X}(l)|\Gamma_V\right) - \log p\left(\mathcal{X}(l)|\Gamma_N\right)\right] \underset{\text{nonvoc}}{\overset{\text{voc}}{\gtrless}} \psi. \quad (14)$$

C. Acoustic Adaptation of the Music Model

The acoustically adapted music model is estimated from the nonvocal parts $\{X(t)\}_{t\notin\mathbf{voc}}$ using a MAP criterion

$$\tilde{\lambda}_m = \arg\max_{\lambda'_m} p\left(\{X(t)\}_{t\notin\mathbf{voc}}\middle|\lambda'_m\right) p\left(\lambda'_m\middle|\Lambda_m\right) \qquad (15)$$

where, following [17], the prior pdf $p(\lambda_m|\Lambda_m)$ is chosen as the product of pdfs of conjugate priors for the model parameters (i.e., normal inverse Wishart priors for covariance matrices, and Dirichlet priors for Gaussian weights). These priors involve a *relevance factor* $\tau \in [0,\infty)$ as a parameter representing the degree of attachment of the adapted model λ_m to the general one Λ_m. This MAP criterion, with such a choice for the priors, can be optimized using the EM algorithm [15] leading to the reestimation formulas which can be found in [17].

[1]Note that the structure of these GMMs is slightly different from that of the GMMs Λ_v and Λ_m used for separation. In particular, the observation vectors are real (not complex), and the mean vectors are not zero.

This way of estimating an acoustically adapted music model calls for both prior knowledge and auxiliary information, thus fitting in the general formalism introduced in the previous section:

- the pretrained model Λ_m, which expresses prior statistical knowledge on the music source and translates into an attachment constraint of the observed source parameters to the general model;
- the segmentation of the mix between vocal and nonvocal parts, where the latter represents auxiliary information indicating when the mix X can be considered as pertaining to the music source only, which can be resorted to for improving the estimation of the music model.

In the general case, these two sources of knowlege and information are combined in the maximization of criterion (15). However, two extreme cases may occur in particular practical situations.

- *Full-Retrain* ($\tau = 0$): The nonvocal parts are in a significant and sufficient quantity to allow a complete reestimation of the music model without resorting to the prior knowledge from the general music model: in that case, the MAP approach degenerates into an ML estimation of the adapted music model.
- *No-Adapt* ($\tau \to \infty$): Very few or even no nonvocal parts at all are detected in the mix X; no auxiliary information is thus available, and therefore the general model constitutes the only source of knowledge that is exploitable to constrain the solution of the separation procedure.

D. Adaptation of Filters and PSD Gains

In this section, an adaptation technique called *adaptation of filters and PSD gains* (Fig. 4) is presented. This technique falls in the proposed adaptation formalism (Section III).

It can be viewed as a *constrained adaptation technique*, which is presented below in a general form. Next, it is explained how such a technique fits in our proposed adaptation formalism. Then, the techniques of *filter adaptation* and *PSD gain adaptation* are introduced separately. Finally, it is shown how these techniques can be assembled to form a joint *filter and PSD gain adaptation*.

1) Constrained Adaptation: Constrained adaptation is based on the assumption that the parameters of each adapted model λ_k belong to some *subset of admissible parameters* $\Xi_k(\Lambda_k)$ which depends on the parameters of the corresponding general model Λ_k. It is supposed as before that the parameters of the model λ_k possess some prior density, which is defined on this subset and depends also on the general model Λ_k. For example, the parameters of the adapted model λ_k can depend on those of the general model Λ_k via some parametric deformation $\mathcal{L}_k$ with free parameters C_k, i.e., $\lambda_k = \mathcal{L}_k(C_k, \Lambda_k)$. The goal of constrained adaptation is to find the free parameters C_v and C_m satisfying the following MAP criterion:

$$(C_v, C_m) = \arg \max_{C'_v, C'_m} p\left(X|\lambda'_v, \lambda'_m\right) p\left(C'_v|\Lambda_v\right) p(C'_m|\Lambda_m)$$
$$\text{subject to} \quad \lambda'_v = \mathcal{L}_v\left(C'_v, \Lambda_v\right) \quad \text{and} \quad \lambda'_m = \mathcal{L}_m\left(C'_m, \Lambda_m\right) \tag{16}$$

where $p(C_k|\Lambda_k)$, $k = v, m$ are the prior pdfs for the free parameters.[2] The adapted models are then obtained as $\lambda_k = \mathcal{L}_k(C_k, \Lambda_k)$, $k = v, m$.

From a strict mathematical point of view, the MAP criterion (16) is different from criterion (12), but from a practical point of view they are similar. Indeed, the additional parametric constraints (i.e., $\lambda_k = \mathcal{L}_k(C_k, \Lambda_k)$) play a similar role to that of the prior pdfs and the EM algorithm (see the Appendix) is still applicable for criterion (16).

2) Filter Adaptation: In our previous work [34], we introduced a constrained adaptation technique consisting in the adaptation of one single filter. With this adaptation, the modeling becomes invariant to any cross-recording variation that can be represented by a global linear filter, for example variation of room acoustics, of some microphone characteristics, etc. The mismatch between the general model Λ_v and the adapted one λ_v can be modeled as a linear filter. In other words, each source modeled by λ_v is considered as a result of filtering with a filter h_v of some other source modeled by Λ_v. The filter h_v is supposed to be unknown, and the goal of the filter adaptation technique is to estimate it.

Let $H_v = \{H_v(f)\}_f$ be the Fourier transform of the impulse response of the filter h_v. We have the following relation between the PSDs of adapted λ_v and general Λ_v models:

$$\sigma^2_{v,i}(f) = |H_v(f)|^2 r^2_{v,i}(f), \quad f = 1, \ldots, F. \tag{17}$$

Introducing the diagonal matrix $\mathcal{H}_v \triangleq \text{diag}[\{\mathcal{H}_v(f)\}_f]$ with $\mathcal{H}_v(f) \triangleq |H_v(f)|^2$ (hereafter, this matrix $\mathcal{H}_v$ will be called *filter*) expression (17) can be rewritten as follows, linking the model λ_v together with the model $\Lambda_v = \{u_{v,i}, R_{v,i}\}_i$:

$$\lambda_v = \mathcal{H}_v \Lambda_v \triangleq \{u_{v,i}, \mathcal{H}_v R_{v,i}\}_i. \tag{18}$$

In the context of constrained adaptation presented in the previous subsection, the filter $\mathcal{H}_v$ plays the role of the free parameters C_v, and $\mathcal{H}_v \Lambda_v$ plays the role of the parametric deformation $\mathcal{L}_v(C_v, \Lambda_v)$. The following criterion, corresponding to criterion (16), is used to estimate the filter $\mathcal{H}_v$:

$$\mathcal{H}_v = \arg \max_{\mathcal{H}'_v} p\left(X \middle| \mathcal{H}'_v \Lambda_v, \tilde{\lambda}_m\right). \tag{19}$$

Note that, since the adaptation is done in two steps (see Fig. 4), the acoustically adapted music model $\tilde{\lambda}_m$ is used in this criterion (19) instead of some general music model Λ_m. Let us also remark that there is no additional constraint on the filter $\mathcal{H}_v$, i.e., there is a noninformative uniform prior ($p(\mathcal{H}_v|\Lambda_v) \propto const$). However, thanks to constraint (18), the adapted model λ_v remains attached to the general one Λ_v. In Appendix D1, we describe in details how to perform the EM algorithm to optimize criterion (19).

Note that the filter adaptation can be considered as a sort of constrained MLLR adaptation. Indeed, the MLLR technique [22], [23] consists in adapting an affine transform of the feature

[2]For the particular constrained adaptation techniques introduced in this paper (cf. Section IV-D2 and IV-D3) noninformative uniform priors are used, i.e., $p(C_k|\Lambda_k) \propto const$. In other words, no particular knowledge is assumed on the values taken by the free parameters C_k.

space, while for filter adaptation, only dilatations and contractions along the axes of the STFT (feature) space are allowed.

3) PSD Gains Adaptation: Each state of a GMM is described by some characteristic spectral pattern (or local PSD) corresponding to some particular sound event, for example, a musical note or chord. The relative mean energies of these sounds events vary between recordings. For example, in one recording, the A note can be played on average louder than the D note, while it can be the opposite for another recording. In order to take into account this energy variation, a positive gain $g_{v,i} > 0$ is associated to each PSD $\{r^2_{v,i}(f)\}_f$ of the model Λ_v. This gain is called *PSD gain* and corresponds to the mean energy of the sound event represented by this PSD. Since each PSD is the diagonal of the corresponding covariance matrix $R_{v,i}$, this matrix is multiplied by the PSD gain $g_{v,i}$. Thus, the PSD gains adaptation technique consists in looking for the adapted model λ_v in the following form:

$$\lambda_v = g_v \bullet \Lambda_v \triangleq \{u_{v,i}, g_{v,i} R_{v,i}\} \tag{20}$$

where $g_v = \{g_{v,i}\}_i$ is a vector of PSD gains and the symbol "$\bullet$" denotes a nonstandard operation used here to distinguish between the application of the PSD gains $(g_v \bullet \Lambda_v)$ and that of the filter (18) $(\mathcal{H}_v \Lambda_v)$.

Comparing to the filter adaptation technique, where the goal is to adapt the energy for each frequency band f, the goal of the PSD gains adaptation is to adapt the energy for each PSD i.

The following explication is very similar to the one given for filter adaptation. The PSD gains g_v play the role of free parameters and the following criterion is used to estimate them

$$g_v = \arg\max_{g'_v} p\left(X \middle| g'_v \bullet \Lambda_v, \tilde{\lambda}_m\right). \tag{21}$$

Again, the EM algorithm can be used to reestimate the gains, as explained in Appendix D2.

4) Joint Filters and PSD Gain Adaptation: This section details how to adapt the filters and PSD gains jointly for both models. The adapted voice and music models are represented in the following form: $\lambda_v = g_v \bullet \mathcal{H}_v \Lambda_v$ and $\lambda_m = g_m \bullet \mathcal{H}_m \tilde{\lambda}_m$, where $\mathcal{H}_m$ and g_m denote, respectively, the filter and the PSD gains of the music model λ_m. The following criterion is used to estimate all these parameters:

$$\begin{aligned}&(\mathcal{H}_v, g_v, \mathcal{H}_m, g_m)\\ &\quad = \arg \max_{\mathcal{H}'_v, g'_v, \mathcal{H}'_m, g'_m} p\left(X \middle| g'_v \bullet \mathcal{H}'_v \Lambda_v, g'_m \bullet \mathcal{H}'_m \tilde{\lambda}_m\right).\end{aligned} \tag{22}$$

The direct application of the EM algorithm (29), (30) to optimize criterion (22) is not possible, since it is difficult to solve the M step (cf. (30) in the Appendix) jointly on the filters $\{\mathcal{H}_v, \mathcal{H}_m\}$ and the PSD gains $\{g_v, g_m\}$ (see Appendix D3).

One solution to this problem would be to use the space-alternating generalized EM (SAGE) algorithm [35], [36] alternating the EM iterations between $\{\mathcal{H}_v, \mathcal{H}_m\}$ and $\{g_v, g_m\}$. However, in contrast to the EM algorithm, this approach requires two EM iterations instead of one to reestimate once all the parameters $\{\mathcal{H}_v, g_v, \mathcal{H}_m, g_m\}$. Thus, the computational complexity doubles.

Analyzing separately the computational complexities of the E and M steps (29), (30), we see that the M step computational complexity is negligible in comparison with that of the E step. Indeed, the complexity of the E step (calculation of the natural statistics expectations) is $O(T \times F \times Q_1 \times Q_2)$ [see (38) and (39)] and that of the M step (parameters update) is $O(F \times (Q_1 + Q_2))$ [see, for example, (42) and (43)]. Thus, in order to avoid doubling the complexity, instead of using the SAGE algorithm, we propose for each iteration to do one E step followed by several M steps alternating between the updates of $\{\mathcal{H}_v, \mathcal{H}_m\}$ and $\{g_v, g_m\}$. Algorithm 2 in the Appendix summarizes this principle.

V. Experiments

Experiments concerning model adaptation in the context of voice/music separation are presented in this section. First, the module for automatic vocal/nonvocal segmentation is evaluated independently from the adaptation block (Fig. 4). Then, the experiments on model adaptation and separation are developed, using a manual vocal/nonvocal segmentation in the first place, and an automatic segmentation in the second place.

A. Automatic Vocal/Nonvocal Segmentation

1) Data Description: The training set for learning Γ_V and Γ_N GMMs, modeling vocal and nonvocal parts, contains 52 popular songs. A set of 22 other songs is used to evaluate the segmentation performance. All recordings are mono, sampled at 11 025 Hz, and manually segmented into vocal and nonvocal parts.

2) Acoustic Parameters: Classical MFCC-based acoustic parameters are chosen for this segmentation task. In particular, the vector of parameters for each frame consists of the first 12 MFCC coefficients [33] and the energy (13 parameters), their first- and second-order derivatives Δ and $\Delta\Delta$ (which represents 39 parameters in total). The MFCC coefficients are obtained from the STFT, which is computed using a half-overlapped 93-ms length Hamming window. Parameters are normalized using cepstral mean subtraction (CMS) and variance normalization (VN) [37] in order to reduce the influence of convolutive and additive noises.

3) Performance Measure: The performance of vocal/nonvocal segmentation is evaluated using detection error tradeoff (DET) curves [38]. For a given segmentation threshold ψ [see (13), (14)], the segmentation performance can be evaluated in terms of two types of errors: the vocal miss error rate (VMER), which is the rate of vocal frames identified as nonvocal and the vocal false alarm rate (VFAR), which is the rate of nonvocal frames identified as vocal.

These error measures are computed comparing the automatic segmentation with a manual one. Frames localized in a 0.5-s interval around a manually marked switch-point are not taken into account for the calculation of VMER and VFAR. This tolerance is justified by the fact that it is difficult to mark accurately the switch-points between vocal and nonvocal parts by hand. The coordinates of each point of a DET curve are the VMER and the VFAR as the segmentation threshold varies.

4) Simulations: A 32-Gaussian GMM Γ_V and a 32-Gaussian GMM Γ_N are learned from the training data using 50 iterations

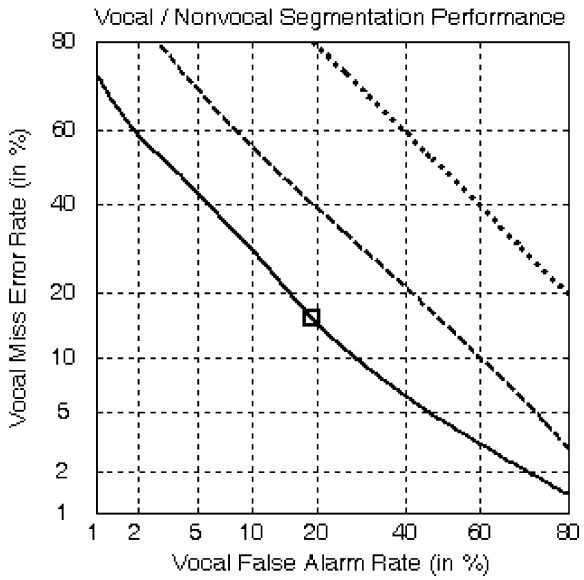

Fig. 5. DET curves for vocal/nonvocal automatic segmentation. Dotted line: random segmentation, EER $= 50\%$. Dashed line: frame-based decision [see (13)], EER $= 29\%$. Solid line: block-based decision [see (14)] with 1-s block length, EER $= 17\%$. Square: operating point chosen for model adaptation ($\psi = 0$, VMER $= 15\%$, VFAR $= 19\%$).

of the EM algorithm,[3] which is initialized by the K-means (or Lloyd) algorithm [39]. Segmentation results are represented on Fig. 5. With the frame-based decision (13), the equal error rate (EER) (i.e., when VMER $=$ VFAR) is 29%. Note that a random segmentation gives an EER of 50%. When the block-based decision (14) with 1-s block length is used, the EER significantly falls down to 17%.

Since one goal of this work is to improve the separation performance, the threshold ψ (corresponding to some operating point on the DET curve) for segmentation system integrated in the model adaptation scheme (Fig. 4) should be chosen on the basis of the separation performance. This issue is addressed in the following section.

Note that, in the choice of the segmentation threshold, there is a tradeoff between purity and quantity of data. Indeed, since the nonvocal parts are used for the music model acoustic adaptation (Fig. 4), from one side the nonvocal parts should be quite pure, or not much disturbed by vocal frames detected by mistake, i.e., the VMER should be low. On the other side, a sufficient quantity of nonvocal frames should be detected correctly in order to have enough data to adapt the music model, i.e., the VFAR should be low.

B. Adaptation and Separation

1) Data Description: The training database for the general voice model Λ_v includes 34 samples of "pure" singing voice from popular music. The general music model Λ_m is trained on 30 samples of popular music free from voice. Each sample is approximately one minute long. The test database contains six popular songs, for which voice and music tracks are available separately. It is therefore possible to evaluate the separation performance by comparing the estimated voice with the original one. The test items are manually segmented into vocal and nonvocal parts (automatic segmentation is also performed in the experiment). All recordings are mono and sampled at 11 025 Hz.

2) Parameters: As for segmentation, the STFT is computed using a half-overlapped 93-ms-length Hamming window.

3) Performance Measure: Separation performance is estimated using the Normalised SDR (NSDR) [34], which measures the improvement of the source to distortion ratio (SDR) [40] in decibels

$$\mathrm{SDR}(\hat{s}_k, s_k) = 10\log_{10}\left[\frac{\langle \hat{s}_k, s_k\rangle^2}{||\hat{s}_k||^2||s_k||^2 - \langle \hat{s}_k, s_k\rangle^2}\right] \quad (23)$$

between the nonprocessed mix x and the estimated source $\hat{s}_k$:

$$\mathrm{NSDR}(\hat{s}_k, s_k, x) = \mathrm{SDR}(\hat{s}_k, s_k) - \mathrm{SDR}(x, s_k). \quad (24)$$

The aim of this normalization is to combine the absolute measure $\mathrm{SDR}(\hat{s}_k, s_k)$ and the "difficulty" of the separation task for processed recording $\mathrm{SDR}(x, s_k)$. This difficulty is expressed as the performance of "inactive separation," i.e., when the mix x itself is taken instead of the estimate $\hat{s}_k$. The higher the NSDR, the better the separation performance.

In the context of audio indexing, we are mainly interested in voice estimation $\hat{s}_v$ (Section IV). Therefore, the separation performance is evaluated using the voice NSDR (i.e., $\mathrm{NSDR}(\hat{s}_v, s_v, x)$) and not the music one. Note, at the same time, that the order of the music NSDR is quite similar to that of the voice NSDR. The overall performance is estimated by averaging the voice NSDRs calculated for all songs from the test database.

4) Simulations: In order to estimate the efficiency of each step in the proposed adaptation scheme, as well as the efficiency of the adaptation of different parameter combinations (filters, PSD gains), the separation experiments are performed with 32-state voice GMM and 32-state music GMM in the following configurations.

1) **General models**: Λ_v and Λ_m are learned from external training data (50 iterations of the EM algorithm, initialized by the K-means algorithm).
2) **Acoustically adapted models**:
   - *Voice model*: As mentioned in Section IV-A, the mix X is segmented into vocal and nonvocal parts. The vocal parts correspond to portions of the signal that include voice, but only an unsignificant part of these portions may contain only voice signals (most of them are composed of voice + music). Therefore, these data are not used in the acoustic adaptation of the voice model. The voice GMM is kept constant, i.e., $\tilde{\lambda}_v = \Lambda_v$, which corresponds to the degenerate case "no-adapt" of Section IV-C.
   - *Music model*: Experiments have been run to determine the optimal relevance factor τ (see Section IV-C) for adapting the music GMM in the MAP framework. For our test data set, the optimal value was observed to be zero, i.e., the "full-retrain" degenerate case of Section IV-C.[4] The EM algorithm run in this context

[3]The numbers of the EM algorithm iterations (here 50) reported hereafter were found suitable for guaranteeing appropriate convergence of the algorithm in each particular implementation.

[4]This situation may arise from the fact that the general music model is not very elaborate and that the number of music-only frames (about 200–500) segmented from the mix is in sufficient quantity for every song.

TABLE I
AVERAGE NSDR ON THE SIX SONGS OF THE TEST DATABASE OBTAINED WITH DIFFERENT MODEL TYPES ($Q_v = Q_m = 32$)

Model type	Voice model	Music model	Segmentation	
			manual	automatic
General models (state of the art)	Λ_v	Λ_m	5.4	
Acoustically adapted models	Λ_v	$\tilde{\lambda}_m$	11.3	9.5
Adapted models (filters, PSD gains)	$\lambda_v = \mathcal{H}_v\Lambda_v$	$\lambda_m = \tilde{\lambda}_m$	12.3	10.2
	$\lambda_v = g_v \bullet \mathcal{H}_v\Lambda_v$	$\lambda_m = \tilde{\lambda}_m$	**12.8**	**10.5**
	$\lambda_v = g_v \bullet \mathcal{H}_v\Lambda_v$	$\lambda_m = g_m \bullet \tilde{\lambda}_m$	12.5	10.4
	$\lambda_v = g_v \bullet \mathcal{H}_v\Lambda_v$	$\lambda_m = g_m \bullet \mathcal{H}_m\tilde{\lambda}_m$	12.2	10.4
Ideal models (empirical bound)	λ_v^{Idl}	λ_m^{Idl}	*15.9*	

was iterated 40 times after initialization by the K-means algorithm.

3) **Filter/Gain Adapted models**: λ_v and λ_m are obtained from the models Λ_v and $\tilde{\lambda}_m$ via an adaptation of the following parameter combinations:[5]
   a) filter-adapted for voice model $\{\mathcal{H}_v\}$;
   b) filter- and gain-adapted for voice model $\{\mathcal{H}_v, g_v\}$;
   c) filter-adapted for voice model and gain-adapted for both models $\{\mathcal{H}_v, g_v, g_m\}$;
   d) filter- and gain-adapted for both models $\{\mathcal{H}_v, g_v, \mathcal{H}_m, g_m\}$.

   (Five iterations of EM algorithm 2 described in Appendix D, initialized as follows: $g_k = [1, 1, \ldots, 1]^T$, $\mathcal{H}_k = I$ for $k = v, m$, where I is the identity matrix.)
4) **Ideal models**: λ_v^{Idl} and λ_m^{Idl} are learned from the separated sources S_v and S_m, which are available for evaluation purposes (40 iterations of the EM algorithm, initialized by the K-means algorithm). The separation performance obtained with these "ideal" models (inaccessible in a real application context) acts as a kind of empirical upper bound for the separation performance, which can be obtained with adapted models.

Since the estimation of the acoustically adapted music model $\tilde{\lambda}_m$ is based on some vocal/nonvocal segmentation, the tests involving this model are performed using both manual and automatic segmentation.

Automatic segmentation is done by a block-based decision system (14) with 1-s block length and with a segmentation threshold $\psi = 0$. These parameters were chosen since they lead to the best separation performance with acoustically adapted models. The segmentation performances of this system are VMER $= 15\%$ and VFAR $= 19\%$ (see Fig. 5).

The average results on the six songs of the test database are summarized in Table I. A main performance improvement is obtained with the acoustic adaptation of the music model $\tilde{\lambda}_m$ from the nonvocal parts. There is a 5.8 and 4.1 dB improvement for manual and automatic segmentations, respectively.

The voice model filter $\mathcal{H}_v$ adaptation increases further the performance by 1.0 and 0.7 dB for the two types of segmentation. An additional adaptation of the PSD gains g_v for this model leads also to a slight performance improvement. Adaptation of the music model parameters (i.e., $\mathcal{H}_m$ and g_m) does not increase the performance any further. This can be explained by the fact that the music model $\tilde{\lambda}_m$ is already quite well adapted by the acoustic adaptation step.

[5]Note that Algorithm 2 is still applicable with slight modifications, when only a part of parameters $\{\mathcal{H}_v, g_v, \mathcal{H}_m, g_m\}$ is adapted. For example, to adapt $\{\mathcal{H}_v, \mathcal{H}_m, g_m\}$, (45) should be skipped.

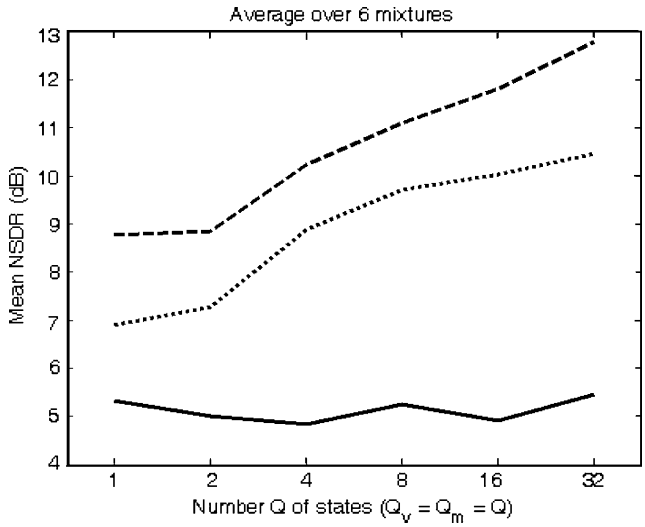

Fig. 6. Average NSDR on the six songs of the test database for different numbers of states $Q = Q_v = Q_m$ and for different types of models. Plain: general models. Dotted: adapted models with automatic segmentation. Dashed: adapted models with manual segmentation.

Altogether, compared with the general models, adaptation improves the separation performance by 7.4 dB with a manual segmentation and still by 5.1 dB when the segmentation is completely automatic. One can note that these results are 3.1 and 5.4 dB below the empirical upper bound obtained using ideal models. It remains a challenge to reduce this gap with improved model adaptation schemes.

The effect of model dimensionality (i.e., number of states $Q = Q_v = Q_m$) on the separation performance is evaluated in the following configurations:

1) general models Λ_v and Λ_m;
2) adapted models $\lambda_v = g_v \bullet \mathcal{H}_v\Lambda_v$ and $\lambda_m = \tilde{\lambda}_m$ (giving the best separation results according to Table I) with manual segmentation;
3) adapted models $\lambda_v = g_v \bullet \mathcal{H}_v\Lambda_v$ and $\lambda_m = \tilde{\lambda}_m$ with automatic segmentation.

The results are represented in Fig. 6. Note that increasing the number of states in the case of general models does not lead to performance improvement, compared with one-state GMMs ($Q_v = Q_m = 1$). A one-state GMM consists of only one PSD, thus for $Q_v = Q_m = 1$ the Wiener filter defined by (8) is a linear filter, which does not vary in time. It was noticed that

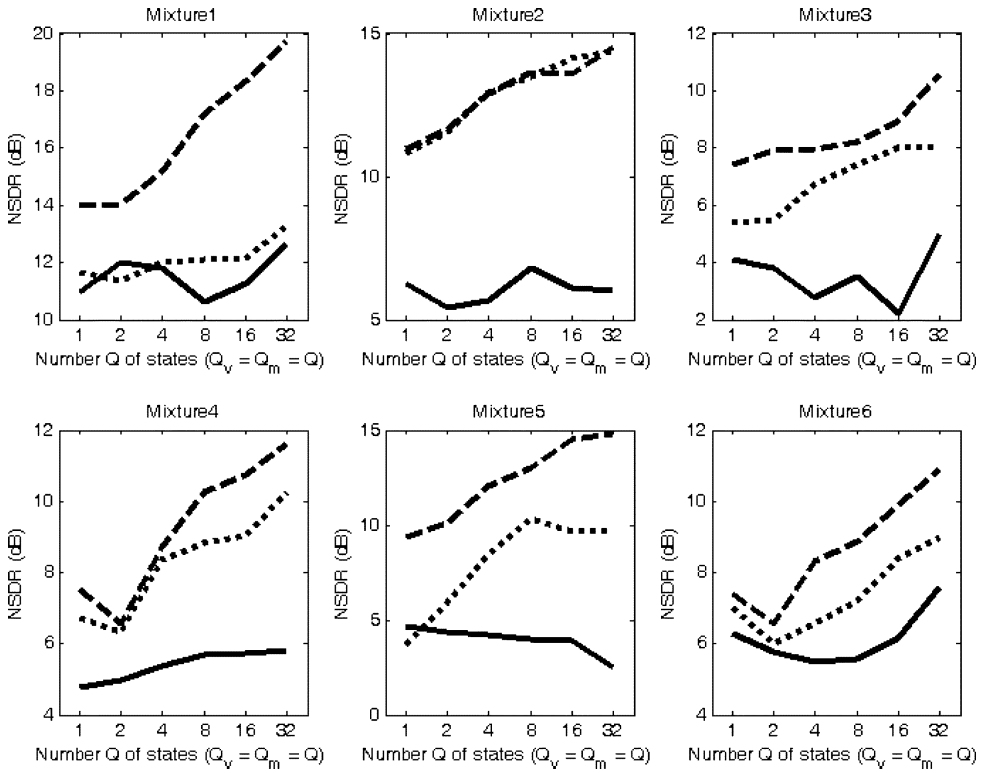

Fig. 7. Detailed NSDR on the six songs of the test database for different numbers of states $Q = Q_v = Q_m$ and for different types of models. Plain: general models. Dotted: adapted models with automatic segmentation. Dashed: adapted models with manual segmentation.

for voice estimation, this is merely a high-pass filter with its cutoff frequency around 300 Hz. Thus, for the voice/music separation task, the general models cannot give better performance than 5-dB NSDR obtained with a simple high-pass filtering, and therefore there is no interest to use general models with several states. Probably, this is because of the problem of weak representativeness of training data for wide sound classes mentioned in Section II-B.

As illustrated by our experiments, the model adaptation allows to overcome these limits. Indeed, with adapted models, the separation performance can be significantly improved by increasing the number of model states, and the added computational complexity pays off. This experiment indicates that for source separation tasks with wide sound classes (such as music), model adaptation is essential.

As can be seen in Fig. 7, a deeper investigation of the behavior of the NSDR for each of the six test songs separately shows a consistent behavior of the proposed adaptation scheme.

Concerning computational complexity, it is worth mentioning that the proposed system needs about 4 h to separate 23 min (total duration of six test songs) using a laptop equipped with Pentium M processor 1.7 GHz, which is quite reasonable.

Note that the problem of voice/music separation in monophonic recordings is a very difficult task which was not much studied ([34], [41], [42]). For this task, we have developed a separation system, which, thanks to model adaptation, has the following advantages.

- Compared with general models, the separation performance is improved by 5 dB.
- The system is completely automatic.
- The computational complexity is quite reasonable (less than ten times RT).
- Experiments were carried out with no special restrictions about music style (while staying in pop/rock songs) nor about the language of the songs.

However, there are also some limitations, which should be mentioned. First, the processed song must contain nonvocal parts of reasonable length in order to have enough data for the acoustic adaptation of music model. Second, the music from nonvocal parts should be quite similar to that from vocal parts. Finally, it is preferable that there is only one singer at a time, i.e., no chorus or back vocals. At first sight, a majority of popular songs verify these assumptions.

VI. Conclusion and Further Work

In the context of probabilistic methods for source separation with a single channel, we have presented a general formalism consisting in *a posteriori* model adaptation. This formalism is introduced in the general form of Bayesian models, and further clarified in terms of a MAP adaptation criterion which can be optimized using the EM algorithm.

To show the relevance of model adaptation in practice, a model adaptation system derived from this formalism has been designed for the difficult task of separating voice from music in popular songs. This system is based on vocal/nonvocal segmentation, on adapting acoustically a music model from the nonvocal parts, and on a final adaptation of voice and music models from the mix using filters and PSD gains adaptation technique.

Compared to general (nonadapted) models, the adaptation allows, in our experiments, to consistently improve the separation performance. It yields on average a 5-dB improvement which bridges half of the gap between the use of general models on the one hand and ideal models on the other hand.

More generally, by formulating the adaptation process in a rather general way, which integrates prior knowledge, structural constraints, and *a posteriori* observations, the work reported in this paper may contribute to the solution of a number of problems, whether they resort to blind, knowledge-based or data-driven source separation.

APPENDIX

In this Appendix, the EM algorithm [15] is applied to optimize the MAP criterion (12). This algorithm is first presented in its general form, then precisions are given in the case of exponential families. Finally, some additional calculations are done for the GMMs studied in this article.

A. EM Algorithm in its General Form

The following notations are introduced with their names given according to the terminology of the EM algorithm [15], [36]:

- X observed data;
- $Z \triangleq \{q_1, S_1, q_2, S_2\}$ complete data (recall that $q_k = \{q_k(t)\}_t$, $k = 1, 2$ denote GMM state sequences);
- $\theta \triangleq \{\lambda_1, \lambda_2\}$ estimated parameters;
- $p(\theta) \triangleq p(\lambda_1|\Lambda_1)p(\lambda_2|\Lambda_2)$ prior pdf

Note that the observed and complete data are chosen in an appropriate way to use EM. Indeed, according to (3), the observed data X are expressed in a unique manner from the complete data Z.

With these new notations, the MAP criterion (12) can be rewritten in a more compact form

$$\theta^{\mathrm{MAP}} = \arg\max_{\theta'} p(X|\theta')p(\theta'). \tag{25}$$

In order to optimize this MAP criterion, the EM algorithm is used. This algorithm is an iterative procedure, and in its general form can be expressed as follows [15], [36]:

$$Q\left(\theta, \theta^{(l)}\right) = \mathbb{E}_Z\left[\log p(Z|\theta)\middle| X, \theta^{(l)}\right] + \log p(\theta) \tag{26}$$

$$\theta^{(l+1)} = \arg\max_{\theta} Q\left(\theta, \theta^{(l)}\right) \tag{27}$$

where $\theta^{(l)}$ denotes the parameters estimated at the lth iteration. The E step (expectation) (26) consists in computing an auxiliary function $Q(\theta, \theta^{(l)})$, and the M step (maximization) (27) consists in estimating the new parameters maximizing this function.

B. EM Algorithm for Exponential Families

The EM algorithm takes a particular form if the families of complete data pdfs $\{p(S_k, q_k|\lambda_k)\}_{\lambda_k}$, $k = 1, 2$ are exponential families, as recalled in Definition 1 below. This is the case for the GMMs (as shown in Appendix C1), as well as for the HMMs. In this paper, we present the EM algorithm for this particular case of exponential families, since we believe that in this form, the algorithm is easier to understand, and its derivation for the GMMs, as well as for the HMMs, becomes very compact and straightforward.

Definition 1: References [15], [36]. The family of pdfs $\{p(V|\eta)\}_\eta$ parameterized by η is called an *exponential family* if $p(V|\eta)$ can be expressed in the following form:

$$p(V|\eta) = \exp\left\{\langle g(\eta), \mathbf{T}(V)\rangle + d(\eta) + h(V)\right\} \tag{28}$$

where $d(\eta), h(V) \in \mathbb{R}$ are scalar functions, $g(\eta), \mathbf{T}(V) \in \mathbb{R}^L$ vector functions, and $\langle\cdot,\cdot\rangle$ denotes scalar product. The function $\mathbf{T}(V)$ is called *natural statistics* for this exponential family.

The natural statistics $\mathbf{T}(V)$ are also *sufficient* [14] for the parameter η. For any sufficient statistics, the following property is fulfilled.

Property 1: If $\mathbf{T}(V)$ is a sufficient statistics for η, then the MAP parameter estimator $\eta^{\mathrm{MAP}} = \arg\max_{\eta'} p(V|\eta')p(\eta')$ must be a function of $\mathbf{T}(V)$.

Here, denoting $\mathbf{T}_k(S_k, q_k)$ the respective natural statistics of $p(S_k, q_k|\lambda_k)$, it can be shown [15], [36] that the EM algorithm (26), (27) can be represented in a form which is easier to understand, to interpret and to use, specifically

$$\mathbf{T}_k^{(l)}(S_k, q_k) = \mathbb{E}_{S_k, q_k}\left[\mathbf{T}_k(S_k, q_k)\middle| X, \lambda_1^{(l)}, \lambda_2^{(l)}\right] \tag{29}$$

$$\lambda_k^{(l+1)} = \mathbf{f}_k\left(\mathbf{T}_k^{(l)}(S_k, q_k)\right) \tag{30}$$

with the functions $\mathbf{f}_k(\mathbf{T}_k(S_k, q_k))$, $k = 1, 2$ defined as solutions of the following complete data MAP criteria

$$\mathbf{f}_k\left(\mathbf{T}_k(S_k, q_k)\right) \triangleq \arg\max_{\lambda_k'} p\left(S_k, q_k|\lambda_k'\right) p\left(\lambda_k'|\Lambda_k\right). \tag{31}$$

The existence of such functions depending only on the natural (sufficient) statistics $\mathbf{T}_k(S_k, q_k)$ is guaranteed by Property 1. Note that the MAP criteria (31) correspond to the MAP criterion (12) assuming the complete data $Z = \{q_1, S_1, q_2, S_2\}$ are observed.

The following simple interpretation can be given to this EM algorithm. If the complete data Z were observed, we would use the complete data MAP criteria (31), and their solutions are $\lambda_k = \mathbf{f}_k(\mathbf{T}_k(S_k, q_k))$. However, since the complete data are not observed, the values of natural statistics $\mathbf{T}_k(S_k, q_k)$ are replaced by their expectations (29) conditionally on the observed data X and the models estimated at the previous iteration. Thus, the E step (29) consists of computing the conditional expectations of the sufficient statistics, and the M step (30) consists of estimating the new model parameters using these expectations.

C. EM Algorithm for GMMs

1) Natural Statistics for GMMs: For the GMMs used throughout this paper, the families of pdfs $\{p(S_k, q_k|\lambda_k)\}_{\lambda_k}$ are exponential families and their natural statistics are

$$\mathbf{T}_k(S_k, q_k) = \left\{\mathbf{t}_{k,i}^0, \left\{\mathbf{t}_{k,i}^2(f)\right\}_f\right\}_i \tag{32}$$

with

$$\mathbf{t}_{k,i}^0 = \sum_t \delta\left(q_k(t), i\right) \tag{33}$$

and

$$\mathbf{t}_{k,i}^2(f) = \sum_t |S_k(t,f)|^2\, \delta\left(q_k(t), i\right) \tag{34}$$

where $\delta(i,j)$ is the Kronecker delta function, which equals to 1 if $i=j$, and equals to 0 otherwise.

Indeed, using the GMM definition (5), the log-likelihood of the complete data $\log p(S_k,q_k|\lambda_k)$ can be expressed as follows:

$$\log p(S_k,q_k|\lambda_k)=\sum_i\left(\left[\log\omega_{k,i}-\sum_f\log\left\{\pi\sigma^2_{k,i}(f)\right\}\right]\mathbf{t}^0_{k,i}-\sum_f\frac{\mathbf{t}^2_{k,i}(f)}{\sigma^2_{k,i}(f)}\right) \quad (35)$$

where the statistics $\mathbf{t}^0_{k,i}$ are $\mathbf{t}^2_{k,i}(f)$ are defined according to (33) and (34). Equation (35) can be rewritten as $p(S_k,q_k|\lambda_k)=\exp\{\langle g_k(\lambda_k),\mathbf{T}_k(S_k,q_k)\rangle\}$, where $g_k(\lambda_k)$ is some vector function, and the statistics $\mathbf{T}_k(S_k,q_k)$ are defined as in (32).

The statistics $\mathbf{t}^0_{k,i}$ count the number of times that state i has been observed, and the statistics $\mathbf{t}^2_{k,i}(f)$ represent the energy of the STFT S_k associated to state i and calculated in the frequency band f.

2) Conditional Expectations of Natural Statistics for GMMs: The conditional expectations (29) of the natural statistics (32)–(34) are calculated using Algorithm 1. Indeed, (36) is analogous to (9), and (37) can be found in the article of Rose *et al.* [27]. The proof for (39) is given in (40) using a shorthand $\xi^{(l)}\triangleq\{X,\lambda_1^{(l)},\lambda_2^{(l)}\}$, and (38) can be proven in a similar way.

Algorithm 1 Calculation of the conditional expectations of natural statistics for S_1 (and similarly for S_2)

1) Compute the weights $\gamma^{(l)}_{i,j}(t)$ satisfying $\sum_{i,j}\gamma^{(l)}_{i,j}(t)=1$ and

$$\gamma^{(l)}_{i,j}(t)\triangleq P\left(q_1(t)=i,q_2(t)=j\middle|X,\lambda_1^{(l)},\lambda_2^{(l)}\right)\propto\omega^{(l)}_{1,i}\omega^{(l)}_{2,j}N_C\left(X(t);\bar{0},\Sigma^{(l)}_{1,i}+\Sigma^{(l)}_{2,j}\right). \quad (36)$$

2) Compute the expected PSD for state $q_1=i$, $q_2=j$

$$\left\langle|S_1(t,f)|^2\right\rangle^{(l)}_{i,j}\triangleq\mathbb{E}_{S_1}\left[|S_1(t,f)|^2\middle|q_1(t)=i,q_2(t)=j\,X,\lambda_1^{(l)},\lambda_2^{(l)}\right]=\frac{\sigma^{2,(l)}_{1,i}(f)\sigma^{2,(l)}_{2,j}(f)}{\sigma^{2,(l)}_{1,i}(f)+\sigma^{2,(l)}_{2,j}(f)}+\left|\frac{\sigma^{2,(l)}_{1,i}(f)}{\sigma^{2,(l)}_{1,i}(f)+\sigma^{2,(l)}_{2,j}(f)}X(t,f)\right|^2. \quad (37)$$

3) Compute the conditional expectation of $\mathbf{t}^0_{1,i}$

$$\mathbf{t}^{0,(l)}_{1,i}\triangleq\mathbb{E}_{S_1,q_1}\left[\mathbf{t}^0_{1,i}\middle|X,\lambda_1^{(l)},\lambda_2^{(l)}\right]=\sum_t\sum_j\gamma^{(l)}_{i,j}(t). \quad (38)$$

4) Compute the conditional expectation of $\mathbf{t}^2_{1,i}(f)$

$$\mathbf{t}^{2,(l)}_{1,i}(f)\triangleq\mathbb{E}_{S_1,q_1}\left[\mathbf{t}^2_{1,i}(f)\middle|X,\lambda_1^{(l)},\lambda_2^{(l)}\right]=\sum_t\sum_j\left\langle|S_1(t,f)|^2\right\rangle^{(l)}_{i,j}\gamma^{(l)}_{i,j}(t) \quad (39)$$

$$\begin{aligned}\mathbf{t}^{2,(l)}_{1,i}(f)&=\mathbb{E}_{S_1,q_1}\left[\sum_t|S_1(t,f)|^2\,\delta\left(q_1(t),i\right)\middle|\xi^{(l)}\right]\\&=\sum_t\mathbb{E}_{S_1,q_1}\left[|S_1(t,f)|^2\,\delta\left(q_1(t),i\right)\middle|\xi^{(l)}\right]\\&=\sum_t\sum_j\mathbb{E}_{S_1,q_1,q_2}\left[|S_1(t,f)|^2\,\delta\left(q_1(t),i\right)\delta(q_2(t),j)\middle|\xi^{(l)}\right]\\&=\sum_t\sum_j\mathbb{E}_{S_1}\left[|S_1(t,f)|^2\middle|q_1(t)=i,q_2(t)=j,\xi^{(l)}\right]\\&\quad\times P\left(q_1(t)=i,q_2(t)=j\middle|\xi^{(l)}\right)\\&\overset{(36),(37)}{=}\sum_t\sum_j\left\langle|S_1(t,f)|^2\right\rangle^{(l)}_{i,j}\gamma^{(l)}_{i,j}(t).\end{aligned} \quad (40)$$

D. EM for Filters and/or PSD Gains Adaptation

We now have tools to express the EM algorithm in the proposed framework for filters and/or PSD gains adaptation. In order to obtain the reestimation formulas using the EM algorithm (29), (30), one should solve the complete data MAP criteria (31) and express their solutions as functions of natural statistics $\mathbf{T}_k(S_k,q_k)$.

1) Filter Adaptation: In the case of filter adaptation (19), these MAP criteria become (finally, there is only one criterion, since only one model is adapted)

$$\mathcal{H}^*_v=\arg\max_{\mathcal{H}'_v}p(S_v,q_v|\mathcal{H}'_v\Lambda_v) \quad (41)$$

Injecting $\mathcal{H}'_v\Lambda_v$ into expression (35) and zeroing the derivative according to $\mathcal{H}'_v$, one can show that the solution of (41) is given by $\mathcal{H}^*_v(f)=(1/T)\sum_i(\mathbf{t}^2_{v,i}(f)/r^2_{v,i}(f))$. Then, replacing the statistics $\mathbf{t}^2_{v,i}(f)$ by their conditional expectations (39), we have the reestimation formula

$$\mathcal{H}^{(l+1)}_v(f)=\frac{1}{T}\sum_{i=1}^{Q_v}\frac{\mathbf{t}^{2,(l)}_{v,i}(f)}{r^2_{v,i}(f)} \quad (42)$$

where the expectations $\mathbf{t}^{2,(l)}_{v,i}(f)$ are calculated using Algorithm 1, conditionally on the models $\lambda^{(l)}_v=\mathcal{H}^{(l)}_v\Lambda_v$ and $\lambda^{(l)}_m=\tilde{\lambda}_m$.

2) PSD Gains Adaptation: The very same reasoning brings the reestimation formula for PSD gains adaptation

$$g^{(l+1)}_{v,i}=\frac{1}{F\cdot\mathbf{t}^{0,(l)}_{v,i}}\sum_{f=1}^{F}\frac{\mathbf{t}^{2,(l)}_{v,i}(f)}{r^2_{v,i}(f)} \quad (43)$$

with the expectations $\mathbf{t}^{0,(l)}_{v,i}$ and $\mathbf{t}^{2,(l)}_{v,i}(f)$ calculated using Algorithm 1, conditionally on the models $\lambda^{(l)}_v=g^{(l)}_v\bullet\Lambda_v$ and $\lambda^{(l)}_m=\tilde{\lambda}_m$.

3) Joint Filters and PSD Gains Adaptation: Doing the same developments for the criterion (22), i.e., putting $g'_v\bullet\mathcal{H}'_v\Lambda_v$ into expression (35) and zeroing the derivatives according to $\mathcal{H}'_v$ and

g'_v, one can show that the filter $\mathcal{H}^*_v$ is expressed via the PSD gains g^*_v and *vice versa*, i.e.,

$$\mathcal{H}^*_v(f) = \frac{1}{T}\sum\nolimits_i \frac{\mathbf{t}^2_{v,i}(f)}{g^*_{v,i} r^2_{v,i}(f)}$$

and

$$g^*_{v,i} = \frac{1}{F \cdot \mathbf{t}^0_{v,i}} \sum\nolimits_f \frac{\mathbf{t}^2_{v,i}(f)}{\mathcal{H}^*_v(f) r^2_{v,i}(f)}.$$

Thus, we decide to look for the solution alternating between these two expressions, which leads to the reestimation formulas (44)–(47) of Algorithm 2.

Algorithm 2 Joint filters and PSD gain adaptation for models $\Lambda_v = \{u_{v,i}, R_{v,i}\}_i$ and $\tilde{\lambda}_m = \{\tilde{\omega}_{m,j}, \tilde{\Sigma}_{m,j}\}_i$

1) **E step**: Compute the expectations $\{\mathbf{t}^{0,(l)}_{v,i}, \{\mathbf{t}^{2,(l)}_{v,i}(f)\}_f\}_i$ and $\{\mathbf{t}^{0,(l)}_{m,j}, \{\mathbf{t}^{2,(l)}_{m,j}(f)\}_f\}_j$ of the natural statistics conditionally on the models $\lambda^{(l)}_v = g^{(l)}_v \bullet \mathcal{H}^{(l)}_v \Lambda_v$ and $\lambda^{(l)}_m = g^{(l)}_m \bullet \mathcal{H}^{(l)}_m \tilde{\lambda}_m$ using Algorithm 1.
2) **M step**: Update the parameters.
   a) Initialize $g^{[0]}_v = g^{(l)}_v$, $g^{[0]}_m = g^{(l)}_m$;
   b) Perform W maximization steps: for $w = 1, 2, \ldots, W$

$$\mathcal{H}^{[w]}_v(f) = \frac{1}{T}\sum_i \frac{\mathbf{t}^{2,(l)}_{v,i}(f)}{g^{[w-1]}_{v,i} r^2_{v,i}(f)} \tag{44}$$

$$g^{[w]}_{v,i} = \frac{1}{F \cdot \mathbf{t}^{0,(l)}_{v,i}} \sum_f \frac{\mathbf{t}^{2,(l)}_{v,i}(f)}{\mathcal{H}^{[w]}_v(f) r^2_{v,i}(f)} \tag{45}$$

$$\mathcal{H}^{[w]}_m(f) = \frac{1}{T}\sum_j \frac{\mathbf{t}^{2,(l)}_{m,j}(f)}{g^{[w-1]}_{m,j} \tilde{\sigma}^2_{m,j}(f)} \tag{46}$$

$$g^{[w]}_{m,j} = \frac{1}{F \cdot \mathbf{t}^{0,(l)}_{m,j}} \sum_f \frac{\mathbf{t}^{2,(l)}_{m,j}(f)}{\mathcal{H}^{[w]}_m(f) \tilde{\sigma}^2_{m,j}(f)} \tag{47}$$

   c) Set $\mathcal{H}^{(l+1)}_v = \mathcal{H}^{[W]}_v$, $g^{(l+1)}_v = g^{[W]}_v$, $\mathcal{H}^{(l+1)}_m = \mathcal{H}^{[W]}_m$, $g^{(l+1)}_m = g^{[W]}_m$

REFERENCES

[1] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.

[2] G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *J. Mach. Learning Res.*, no. 4, pp. 1365–1392, 2003.

[3] M. Reyes-Gomez, D. Ellis, and N. Jojic, "Multiband audio modeling for single-channel acoustic source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc. (ICASSP'04)*, May 2004, vol. 5, pp. 641–644.

[4] B. Pearlmutter and A. Zador, "Monaural source separation using spectral cues," in *Proc. 5th Int. Conf. Ind. Compon. Anal. (ICA'04)*, 2004, pp. 478–485.

[5] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2001, vol. 13, pp. 793–799.

[6] L. Benaroya and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'03)*, Nara, Japan, Apr. 2003, pp. 957–961.

[7] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'04)*, 2004, vol. 2, pp. 817–820.

[8] G. Peeters and X. Rodet, "SINOLA: A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum," in *Proc. Int. Comput. Music Conf. (ICMC'99)*, Oct. 1999, pp. 153–156.

[9] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, Orlando, FL, May 2002, vol. 3, pp. 3049–3052.

[10] N. H. Pontoppidan and M. Dyrholm, "Fast monaural separation of speech," in *Proc. 23rd Conf. Signal Process. Audio Recording Reproduction Audio Eng. Soc. (AES)* , 2003.

[11] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'04)*, Granada, Spain, Sep. 2004, pp. 327–334.

[12] T. Beierholm, B. D. Pedersen, and O. Winther, "Low complexity Bayesian single channel source separation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'04)*, 2004, vol. 5, pp. 529–532.

[13] D. Ellis and R. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, Toulouse, France, May 2006, vol. 5, pp. 957–960.

[14] S. M. Kay, *Fundamentals of Statistical Signal Processing, Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[16] H. Attias, "New EM algorithms for source separation and deconvolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, vol. 5, pp. 297–300.

[17] J. Gauvain and C. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[18] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, vol. 88, no. 8, pp. 1241–1269, Aug. 2000.

[19] A. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, no. 10, pp. 19–41, 2000.

[20] K. P. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," Ph.D. dissertation, Univ. California Berkeley, Berkeley, CA, Jul. 2002.

[21] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Learning in Graphical Models*, vol. 37, no. 2, pp. 183–233, 1999.

[22] C. Leggetter and P. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *ARPA Spoken Lang. Technol. Workshop*, 1995, pp. 104–109.

[23] M. Gales, D. Pye, and P. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP'96)*, Philadelphia, PA, 1996, vol. 3, pp. 1832–1835.

[24] K. Shinoda and C.-H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE Workshop Speech Recognition Understanding*, Santa Barbara, CA, Dec. 1997, pp. 381–388.

[25] K.-T. Chen, W.-W. Liau, H.-M. Wang, and L.-S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP'00)*, Beijing, China, Oct. 2000, pp. 742–745.

[26] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[27] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 245–257, Apr. 1994.

[28] W.-H. Tsai, D. Rogers, and H.-M. Wang, "Blind clustering of popular music recordings based on singer voice characteristics," *Comput. Music J.*, vol. 28, no. 3, pp. 68–78, 2004.

[29] A. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'01)*, 2001, pp. 119–122.
[30] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR'02)*, Oct. 2002, pp. 164–169.
[31] T. L. Nwe, A. Shenoy, and Y. Wang, "Singing voice detection in popular music," in *Proc. ACM Multimedia Conf.*, New York, Oct. 2004, pp. 324–327.
[32] W. H. Tsai and H. M. Wang, "Automatic detection and tracking of target singer in multi-singer music recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'04)*, Montreal, QC, Canada, 2004, vol. 4, pp. 221–224.
[33] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 525–532, Sep. 1999.
[34] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'05)*, Mohonk, NY, Oct. 2005, pp. 90–93.
[35] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2664–2677, Oct. 1994.
[36] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.
[37] C.-P. Chen, J. Bilmes, and K. Kirchhoff, "Low-resource noise-robust feature post-processing on aurora 2.0," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP'02)*, 2002, pp. 2445–2448.
[38] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. Speech Commun. Technol. (EuroSpeech'97)*, 1997, pp. 1895–1898.
[39] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[40] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'03)*, Apr. 2003, pp. 763–768.
[41] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR'05)*, 2005, pp. 337–344.
[42] Y. Li and D. L. Wang, "Singing voice separation from monaural recordings," in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR'06)*, 2006, pp. 176–179.

Alexey Ozerov received the M.Sc. degree in mathematics from the Saint Petersburg State University, Saint Petersburg, Russia, in 1999, the M.Sc. degree in applied mathematics from the University of Bordeaux 1, Bordeaux, France, in 2003, and the Ph.D. degree in signal processing from the University of Rennes 1, Rennes, France, in 2006.

He was with Orange Labs, Cesson Sévigné, France, and the IRISA, Rennes, from 2003 to 2006 while working towards the Ph.D. degree. From 1999 to 2002, he was an R&D software engineer with Terayon Communication Systems, first in Saint Petersburg and then in Prague, Czech Republic. He is currently a Postdoctoral Researcher in the Sound and Image Processing (SIP) Laboratory, KTH (Royal Institute of Technology), Stockholm, Sweden. His research interests include speech recognition, audio source separation, and source coding.

Pierrick Philippe received the Ph.D. in signal processing from the University of Paris, Orsay, France, in 1995.

Before joining Orange Labs, Cesson Sévigné, France, he was with Envivio (2001–2002) and TDF (1997–2001) where his activities were focused on audio signal processing, especially low bit-rate coding. Before that, he spent two years at Innovason, where he developed DSP effects and algorithms for digital mixing desks. He is now a Senior Audio Specialist at Orange Labs, developing audio algorithms especially related to standards. He is an active member of the MPEG audio subgroup. His main interests are signal processing, including low bit-rate audio coding, and sound analysis and processing.

Frédéric Bimbot received the B.A. degree in linguistics from Sorbonne Nouvelle University, Paris, France, in 1987, the telecommunication engineer degree from ENST, Paris, in 1985, and the Ph.D. degree in signal processing in 1988.

In 1990, he joined the French National Center for Scientific Research (CNRS) as a Permanent Researcher. He was with ENST for seven years and then moved to IRISA (CNRS and INRIA), Rennes, France. He also repeatedly visited AT&T Bell Laboratories between 1990 and 1999. He has been involved in several European projects: SPRINT (speech recognition using neural networks), SAM-A (assessment methodology), and DiVAN (audio indexing). He has also been the work-package Manager of research activities on speaker verification, in the projects CAVE, PICASSO, and BANCA. His research is focused on audio signal analysis, speech modeling, speaker characterization and verification, speech system assessment methodology, and audio source separation. He is heading the METISS Research Group at IRISA, dedicated to selected topics in speech and audio processing.

Dr. Bimbot was Chairman of the Groupe Francophone de la Communication Parlée (now AFCP) from 1996 to 2000 and from 1998 to 2003, a member of the International Speech Communication Association Board (ISCA), formerly known as ESCA.

Rémi Gribonval graduated from École Normale Supérieure, Paris, France in 1997. He received the Ph.D. degree in applied mathematics from the University of Paris-IX Dauphine, Paris, France, in 1999.

From 1999 to 2001, he was a Visiting Scholar at the Industrial Mathematics Institute (IMI), Department of Mathematics, University of South Carolina, Columbia. He is currently a Research Associate with the French National Center for Computer Science and Control (INRIA) at IRISA, Rennes, France. His research interests are in adaptive techniques for the representation and classification of audio signals with redundant systems, with a particular emphasis in blind audio source separation.